2011 11th IEEE International Conference on Data Mining

Mining Dominant Patterns in the Sky

Arnaud Soulet* , Chedy Raïssi†, Marc Plantevit ‡ and Bruno Crémilleux§
* *Université François Rabelais de Tours, LI, EA 2101, F-41029, France*
†*INRIA Nancy Grand-Est, France*
‡*Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France*
§*Université de Caen Basse-Normandie, CNRS, GREYC UMR6072, F-14032, France*

***Abstract*—Pattern discovery is at the core of numerous data mining tasks. Although many methods focus on efficiency in pattern mining, they still suffer from the problem of choosing a threshold that influences the final extraction result. The goal of our study is to make the results of pattern mining useful from a user-preference point of view. To this end, we integrate into the pattern discovery process the idea of skyline queries in order to mine *skyline patterns* in a threshold-free manner. Because the skyline patterns satisfy a formal property of dominations, they not only have a global interest but also have semantics that are easily understood by the user. In this work, we first establish theoretical relationships between pattern condensed representations and skyline pattern mining. We also show that it is possible to compute automatically a subset of measures involved in the user query which allows the patterns to be condensed and thus facilitates the computation of the skyline patterns. This forms the basis for a novel approach to mining skyline patterns. We illustrate the efficiency of our approach over several data sets including a use case from chemoinformatics and show that small sets of dominant patterns are produced under various measures.**

***Keywords*-Skyline analysis, Pattern mining, user-preferences.**

I. Introduction

The process of extracting useful patterns from data, called *pattern mining*, is an important tool for data analysis and has been used in a wide range of applications and domains such as bioinformatics [1] or chemoinformatics [2]. Since the pioneering works of Agrawal *et al.* [3], Mannila *et al*. [4], a large amount of work has been developed and many pattern extraction problems are now identified and understood from both theoretical and computational perspectives.

Most existing pattern mining approaches enumerate patterns with respect to a given set of constraints that range from extremely simple to very complex. For instance, given a transaction database, a well-known *"easy"* pattern mining task is to enumerate all itemsets (i.e., sets of items) that appear in at least s transactions. Another mining approach is to extract from a given graph database all subgraphs that have a diameter larger than l, connectivity higher than c, and where each vertex has a degree bounded by d. So far, the community has made great efforts on sophisticated algorithms pushing the constraints deep into the mining process [5]. But it has paid less attention to how to define constraints. In practice, many constraints entail choosing of threshold values such as the well-used minimal frequency. This notion of *"thresholding"* has serious drawbacks. Unless specific domain knowledge is available, the choice is often arbitrary and may lead to a very large number of extracted patterns which can reduce the success of any subsequent data analysis. This drawback is obviously even deeper when several thresholds are needed and have to be combined. A second drawback is the *stringent enumeration aspect*: a pattern is either above or below the thresholds. What about patterns that respect only some thresholds? With this paradigm it is very difficult to apply *subtle selection* mechanisms. There are very few works such as [6] which propose to introduce a softness criterion into the mining process. Other studies blend user preferences in the mining task in order to limit the number of extracted patterns such as the *top-k* patterns [7], [8]. By associating each pattern with a *rank score*, this approach returns an ordered list of the k patterns with the highest score to the user. However, combining several measures to be reflected in a single scoring function is difficult and the performance of top-k approaches are often sensitive to the size of the datasets and to the threshold value, k.

In this work, we focus on making the results of pattern mining *useful from a user-preference point of view*. To this end, we integrate into the pattern discovery process the idea of skyline queries [9] in order to mine *skyline patterns* in a threshold-free manner. Such queries have attracted considerable attention due to their importance in multi-criteria decision making. Briefly speaking, in a multidimensional space where a preference is defined for each dimension, a point a dominates another point b if a is better (i.e., more preferred) than b in at least one dimension, and a is not worse than b on every other dimension. For example, a user selecting a set of patterns may prefer a pattern with a low frequency, short length and a high confidence. In this case, we say that pattern a *dominates* another pattern b if $a.frequency \leq b.frequency$, $a.length \leq b.length$, $a.confidence \geq b.confidence$, where at least one strict inequality holds. Given a set of patterns, the skyline set contains the patterns that are not dominated by any other patterns.

We claim that skyline pattern mining is interesting for several reasons: first, skyline processing does not require any threshold selection or ranking function. Second, the

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.100

formal property of domination satisfied by the skyline patterns gives to the patterns a global interest with semantics easily understood by the user. However, while this notion of skylines has been extensively developed and researched for database applications, it has remained unused for data mining purposes except for a single work on extracting skyline graphs that maximize two measures: the number of vertices and the edge connectivity [10].

Mining skyline patterns, or skypatterns, can be done in a brute-force manner: i.e., mine all patterns in a first step, then run domination tests with respect to the user preferences and finally output the skyline patterns. However, this naive approach is not feasible in practice as the collection of patterns is often too big to be manageable. Obviously, constraints might be introduced to limit the size of the collection but the consistency of the result may be lost (i.e., some skypatterns may not be produced) and the thresholding problem would remain. A key idea of our work is to take benefit of theoretical relationships between pattern condensed representations and skypatterns. These results improve skypattern extraction and we propose, as a main contribution, an efficient approach which only takes as an input the data set and the measures expressing the user preferences and returns skypatterns. To the best of our knowledge, this is *the first work to study theoretically and empirically the feasibility of the skyline pattern mining in a fully generic way* (i.e., with application to various types of patterns).

The paper is organized as follows. Section II reviews some related work. Section III introduces basic definitions and a formal problem statement. The generic framework of skypattern queries is detailed in Section IV. We report an experimental study on several datasets and a case study from the chemoinformatics domain in section V. We conclude in Section VI.

II. Related work

The notion of dominance that we introduced above (see Section III for a formal definition) is at the core of the skyline processing. In this paradigm, the retrieved data points are the ones that are not dominated by any other point in the analysis space. These skyline points can be viewed as *compromise points* with respect to a given set of criteria. Skyline computation is strongly related to mathematical and microeconomics problems such as maximum vectors [11], Pareto set [12] and multi-objective optimization [13]. Since its rediscovery within the database community by Börzsönyi *et al.* [9], many methods have been developed for answering skyline queries that can handle various constraints in different computational environments. Another aspect of preference-based processing is the *top-k* procedure [7], [8]. A ranking function f_r is applied to patterns, and the k best patterns with the highest score with respect to f_r are returned. As previously mentioned, this approach suffers from limitations. The choice of k is not trivial (i.e., the *horizon* problem): a low value may miss useful patterns and a too high value introduces redundancy within the produced patterns (i.e., highly similar patterns). This limitation is the main motivation for the most informative patterns (MIP) that have been recently proposed in [14]. MIPs can be seen as patterns that *locally dominate* other patterns according to a scoring function. This approach shares a similar spirit to our work as it also limits the number of enumerated patterns to a more manageable level. However, in contrast to our study, work on MIPs includes a notion of dominance that is only *local and specific* to subsets of patterns.

One of the earliest findings in the data mining community is that a mining process usually produces large collections of patterns. Many researchers have proposed methods to reduce the size of the output: the constraint-based pattern mining framework [15], the condensed representations [16] and the compression of the dataset by exploiting Minimum Description Length Principle [17], to name a few. A general observation is that patterns represent fragmented knowledge, and often there is no clear view of how the pieces of the puzzle interact and combine to produce a global model. Recent approaches have therefore used schemes such as pattern teams [18], constraint-based pattern set mining [19] and pattern selections [20] that aim to minimize the redundancy and the number of patterns. The common theme in these studies is to select patterns from the initial large set of patterns on the basis of their usefulness in a given context. Often, these methods focus on optimizing a global measure on the discovered pattern set and neglect the relationships between patterns. Moreover, these approaches suffer from a lack of flexibility to express the queries requested by the analyst. For each method, the user has to understand its semantics and express queries satisfying its algorithmic properties and constraints. In addition, some studies take advantage of closed patterns (according to the support measure) to maximize a specific measure such as growth rate for emerging patterns [21] and area for tiling [22], [23].

III. Problem Formulation and Preliminary Definitions

Our study is interesting for several reasons. Firstly, by carefully selecting patterns that are *"the best available"* for a given set of preferences we reduce significantly the output and limit the *"pattern explosion"* curse. The user is *guaranteed* that only the most significant patterns are present in the final result based on his criteria. Secondly, our approach is *parameter-free*. No thresholds are required (solely optional, depending on the analyst needs), and only the preferences and the data set are given as an input.

A. Preliminary definitions

Although the problem can be formulated for any kind of pattern, for simplicity, we will illustrate our definitions using

Table I: Example of a toy data set and measures

(a) A toy data set $\mathcal{D}$

Tid	Items
t_1	$A\ B\ C\ D\ E\ F$
t_2	$A\ B\ C\ D\ E\ F$
t_3	$A\ B$
t_4	D
t_5	$A\ C$
t_6	E

Items	val
A	10
B	55
C	70
D	30
E	15
F	25

(b) Some measures of $\mathcal{M}$

$Name$	Definition
$area$	$X \mapsto freq(X) \times length(X)$
$mean$	$X \mapsto \frac{min(X.val)+max(X.val))}{2}$
$bond$	$X \mapsto \frac{freq(X)}{freq_\vee(X)}$
$aconf$	$X \mapsto \frac{freq(X)}{max(X.freq)}$
gr_1	$X \mapsto \frac{\lvert\mathcal{D}_2\rvert}{\lvert\mathcal{D}_1\rvert} \times \frac{freq(X,\mathcal{D}_1)}{freq(X,\mathcal{D}_2)}$

Table II: A subset of the primitive-based measures

Measure $m \in \mathcal{M}$	Primitive(s)	Operand(s)
$m_1 \theta m_2$	$\theta \in \{+,-,\times,/\}$	$(m_1, m_2) \in \mathcal{M}^2$
$\theta(s)$	$\theta \in \{freq, freq_\vee, length\}$	$s \in \mathcal{S}$
$\theta(s.val)$	$\theta \in \{sum, max, min\}$	$s \in \mathcal{S}$
constant $r \in \Re^+$	-	-
Syntactic expression $s \in \mathcal{S}$	**Primitive(s)**	**Operand(s)**
$s_1 \theta s_2$	$\theta \in \{\cup, \cap, \setminus\}$	$(s_1, s_2) \in \mathcal{S}^2$
$\theta(s)$	$\theta \in \{f, g\}$	$s \in \mathcal{S}$
variable $X \in \mathcal{L}$	-	-
constant $l \in \mathcal{L}$	-	-

itemset patterns. Section IV discusses the computational and theoretical aspects associated with the problem when extracting other patterns. Let $\mathcal{I}$ be a set of distinct literals called *items*, an itemset (or pattern) corresponds to a non-null subset of $\mathcal{I}$. These patterns are gathered together in the language $\mathcal{L}$: $\mathcal{L} = 2^{\mathcal{I}} \setminus \emptyset$. A transactional dataset is a multi-set of patterns of $\mathcal{L}$. Each pattern, named *transaction*, is a database entry. Table I(a) presents a transactional dataset $\mathcal{D}$ where 6 transactions denoted by $t_1, \ldots, t_6$ are described by 6 items denoted by $A, \ldots, F$.

All the measures discussed in this study are based on the set of *primitive-based measures* $\mathcal{M}$ that were first defined in the context of constraint-based pattern mining [24]. Table II presents general definitions of measures and Table I(b) gives some specific examples. As presented in [24], $\mathcal{M}$ defines a very large set of interesting measures.

In addition to the classical operators of $\Re^+$ and $\mathcal{L}$, the function $freq$ denotes the frequency of a pattern, and $length$ its cardinality. The disjunctive support is $freq_\vee(X) = |\{t \in \mathcal{D} | \exists i \in X : i \in t\}|$. Given a function $val : \mathcal{I} \to \Re^+$, we extend it to a pattern X and note $X.val$ the multiset $\{val(i) | i \in X\}$. This kind of function is used with the usual SQL-like primitives sum, min and max. For instance, $sum(X.val)$ is the sum of val for each item of X. Finally, f is the intensive function i.e. $f(T) = \{i \in \mathcal{I} | \forall t \in T,\ i \in t\}$, and g is the extensive function i.e. $g(X) = \{t \in Tid | X \subseteq t\}$.

Definition 1 (Domination): Given a set of measures $M \subseteq \mathcal{M}$, a pattern X *dominates* another pattern Y with respect to M, denoted by $X \succ_M Y$, iff for any measure $m \in M$, $m(X) \geq m(Y)$ and there exists $m \in M$ such that $m(X) > m(Y)$. Two patterns X and Y are said to be *indistinct* with respect to M, denoted by $X =_M Y$, iff $m(X)$ equals to $m(Y)$ for any measure $m \in M$ (if $M = \emptyset$, then $X =_\emptyset Y$). Finally, $X \succeq_M Y$ denotes that $(X \succ_M Y) \vee (X =_M Y)$.

Consider our running example using the data set $\mathcal{D}$ in Table I and suppose that $M = \{freq, area\}$, then the pattern $ABCDEF$ dominates ABC because $freq(ABC) = freq(ABCDEF) = 2$ and $area(ABCDEF) > area(ABC)$. Notice in this case that $ABCDEF$ is indistinct to ABC with respect to $\{freq\}$. Similarly, suppose that $M = \{freq, mean, length\}$, the pattern AC dominates AB because $freq(AC) = freq(AB) = 3$, $|AB| = |AC| = 2$ and $mean(AC) > mean(AB)$.

B. The skypattern mining problem

Given a set of measures M, if a pattern is dominated by another, according to all measures of M, it is irrelevant and must be discarded in the output. The notion of *skyline pattern* formalizes this intuition.

Definition 2 (Skypattern operator): Given a pattern set $P \subseteq \mathcal{L}$ and a set of measures $M \subseteq \mathcal{M}$, a skypattern of P with respect to M is a pattern not dominated in P with respect to M. The skypattern operator $\mathcal{S}ky(P, M)$ returns all the skypatterns of P with respect to M:

$$\mathcal{S}ky(P, M) = \{X \in P |\ \nexists Y \in P : Y \succ_M X\}$$

Given a set of measures $M \subseteq \mathcal{M}$, the *skypattern mining problem is thus to evaluate the query* $\mathcal{S}ky(\mathcal{L}, M)$. For instance, from the toy data set in Table I, $\mathcal{S}ky(\mathcal{L}, \{freq, length\}) = \{ABCDEF,\ AB, AC, A\}$, as illustrated in Fig. 1.

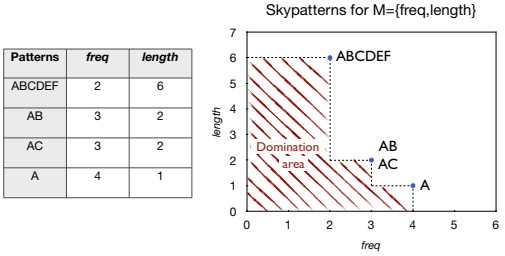


Figure 1: Example of skypattern for a given set of measures

In general, the skypattern mining problem is challenging because of the very high number of candidate patterns (i.e $|\mathcal{L}|$). Indeed, a naive enumeration of $\mathcal{L}$ is not feasible. For example, with 1000 items a naive skypattern approach will need to compute $(2^{1000} - 1) \times |M|$ measures and then compare them. A less naive approach based on heuristics (such as the anti-monotonicity of some measures) may give some results. However, the performance will be closely tied to the underlying properties of the data sets. For instance, in the case of the frequency measure, the density

of the data set plays a major role in the performance and some algorithms are not able to extract frequent patterns at very low thresholds. Nevertheless, considering the following property sheds new insights into an efficient computation of skypattern queries.

Property 1: Given a set of measures $M \subseteq \mathcal{M}$, $\mathcal{S}ky(\mathcal{L}, M)$ equals to $\mathcal{S}ky(P, M)$ for any pattern set P containing $\mathcal{S}ky(\mathcal{L}, M)$,

$$(\forall P \subseteq \mathcal{L})(\mathcal{S}ky(\mathcal{L}, M) \subseteq P \Rightarrow \mathcal{S}ky(\mathcal{L}, M) = \mathcal{S}ky(P, M))$$

As $\mathcal{S}ky(\mathcal{L}, M) \subseteq P \subseteq \mathcal{L}$ and $|P| \leq |\mathcal{L}|$, we argue that evaluating $\mathcal{S}ky(P, M)$ is significantly less costly than evaluating $\mathcal{S}ky(\mathcal{L}, M)$ since the cost of $\mathcal{S}ky(x, M)$ generally decreases with the cardinality of x. Consequently, we aim to reduce the cost of evaluating $\mathcal{S}ky(P, M)$ by finding a small but relevant set P (i.e. that includes $\mathcal{S}ky(\mathcal{L}, M)$) by means of pattern condensed representations. However, this is not an easy task. A direct approach would be to compute a concise representation for each measure $m \in M$, but this is generally not possible because some measures, such as area or length, are *simply not condensable*. Therefore, our problem can be reformulated as following: *given a set of measures M, how can one identify a smaller set of measures M' which allows for the computation of a concise representation on the patterns? In addition, how to use this set of measures to extract efficiently the skypatterns without redundancies?* We address this problem in Section IV.

IV. Reformulating Skypattern Queries

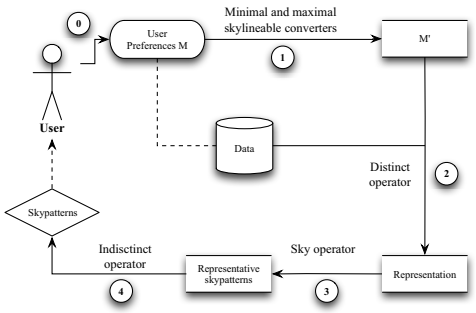


Figure 2: Overview of Aetheris.

In an effort to clarify our methodology, we illustrate in Figure 2 the different processes of our approach called Aetheris. In a first step, and after the user's preferences selection, Aetheris automatically identify a smaller set of measures M' which allows for the computation of a concise representation on the patterns using *converters*. Because of redundancies that may appear in skypatterns, the second step computes a *representative* (i.e., compressed) set of skypatterns. The end-user can either output this compressed representation or the entire list of skypatterns as a final step depending on the application needs. Our methodology revolves around the simple idea that to be able to extract and analyze efficiently skypatterns, one needs to be able to *compress* the patterns that will be used as an input to the skyline operator and then to do a second compression task over the final output (i.e., the skypatterns).

A. Skylineability of a set of measures

Given some specific measures, it is sometimes easy to point out patterns that are excellent skyline candidates. For instance, let us consider patterns from $\mathcal{D}$ that maximize the cardinality. As the cardinality $length(X)$ *strictly* increases with X, the skypattern query $\mathcal{S}ky(\mathcal{L}, \{length(X)\})$ can be defined as a subset of the maximal patterns of $\mathcal{L}$ occurring in $\mathcal{D}$. Unfortunately, this property doesn't hold for other measures such as the frequency (which is only *weakly* decreasing) and the area (which is not monotonic). However, one can notice that the area strictly increases with X when the frequency remains constant. Such a function is said to be *maximally $\{freq\}$-skylineable*.

Definition 3 (Skylineability): Given a set of measures $M' \subseteq \mathcal{M}$, a set of measures M is said to be minimally (respectively maximally) M'-skylineable iff for any patterns $X =_{M'} Y$ such that $X \subset Y$ (respectively $X \supset Y$), one has $X \succeq_M Y$.

Definition 4 (Strict skylineability): Given a set of measures $M' \subseteq \mathcal{M}$ and a set of measures M, if $X \succ_M Y$ for any patterns $X =_{M'} Y$ such that $X \subset Y$ (respectively $X \supset Y$), then M is said to be *strictly* minimally (respectively maximally) M'-skylineable.

From the previous definitions, given a set of measures M which is maximally M'-skylineable, if $X =_{M'} Y$ and $X \supset Y$, it is clear that X cannot be dominated by Y on M. For instance, $M = \{freq, area\}$ is strictly maximally $\{freq\}$-skylineable because $area(X)$ strictly increases with the cardinality of X (when the frequency remains constant). Therefore, in our example, $B =_{freq} AB$ and we can directly deduce that $AB \succ_M B$. Notice that $\{freq\}$ is (weakly) maximally (or minimally) $\{freq\}$-skylineable and that $\{length(X)\}$ is strictly maximally $\emptyset$-skylineable. Next subsections will justifiy the notion of *minimal/maximal* in M'-skylineability by clearly refering to the minimal/maximal patterns of equivalence classes adequate to M'.

Property 2: Any set of measures M is minimally and maximally M-skylineable.

Property 2 is a very important result as it means that *a set of measures is always skylineable*. Obviously, for a set of measures M, the smaller[1] M', the stronger its M'-skylineability. For instance, $\{freq\}$-skylineability is more interesting than $\{freq, area\}$-skylineability because $area$ is not a condensable function: there is no pair of distinct patterns X and Y such that $X =_{\{freq,area\}} Y$. How to choose automatically a subset M' is discussed next.

[1] In the sense of cardinality.

B. Minimal and maximal skylineable converters

Let us first illustrate the general intuition behind an automatic selection technique. Let $M = \{freq\}$ be a set of measures, X and Y be two patterns such that $X \subseteq Y$. Obviously, $M = \{freq\}$ is minimally $\emptyset$-skylineable because $freq$ decreases and $X \succeq_M Y$. Conversely, $M = \{freq\}$ is not *maximally* $\emptyset$-skylineable, but is maximally $\{freq\}$-skylineable. Indeed, if $X =_{\{freq\}} Y$ (i.e., X and Y have the same frequency), then $X \succeq_{\{freq\}} Y$. More generally, any primitive p that is part of the measure m that hinders the M'-skylineability of m, has to be added to M'. We generalize this approach to any primitive-based measure. For this purpose, we define two operators denoted $\underline{c}$ and $\overline{c}$ (see Table III).

Table III: The definition of the minimal and maximal skylineable converters: $\underline{c}$ and $\overline{c}$

Expr. e	Primitive(s)	$\underline{c}(e)$	$\overline{c}(e)$
$e_1 \theta e_2$	$\theta \in \{+, \times, \cup\}$	$\underline{c}(e_1) \cup \underline{c}(e_2)$	$\overline{c}(e_1) \cup \overline{c}(e_2)$
$e_1 \theta e_2$	$\theta \in \{-, /, \cap\}$	$\underline{c}(e_1) \cup \overline{c}(e_2)$	$\overline{c}(e_1) \cup \underline{c}(e_2)$
constant	-	$\emptyset$	$\emptyset$
$d(X)$	$d \in \{freq, min, g\}$	$\emptyset$	$\{d(X)\}$
$i(X)$	$i \in \{length, max, sum, freq_\vee, f\}$	$\{i(X)\}$	$\emptyset$
$d(e_1)$	$d \in \{freq, min, g\}$	$\overline{c}(e_1)$	$\underline{c}(e_1)$
$i(e_1)$	$i \in \{length, max, sum, freq_\vee, f\}$	$\underline{c}(e_1)$	$\overline{c}(e_1)$

Given a primitive-based measure $m \in \mathcal{M}$, the minimal skylineable converter returns a set of measures $M' = \underline{c}(m)$ guaranteeing that for any pattern $X \subset Y$, if $X =_{M'} Y$ then $m(X) \geq m(Y)$. In other words, X dominates Y with respect to m. Dually, the maximal converter $\overline{c}$ guarantees that $m(X) \leq m(Y)$ for any pattern $X \subset Y$ such that $X =_{\overline{c}(m)} Y$.

Let us illustrate $\underline{c}$ and $\overline{c}$ on the area measure. The area is defined as a product of the frequency and length. Thus, we report to the first definition in Table III. $\underline{c}(area) = \underline{c}(freq(X)) \cup \underline{c}(length(X)) = \emptyset \cup \{length(X)\} = \{length(X)\}$. Symmetrically, $\overline{c}(area) = \overline{c}(freq(X)) \cup \overline{c}(length(X)) = \{freq(X)\} \cup \emptyset = \{freq(X)\}$. The skylineable converters enable us to automatically find optimization techniques already known for specific measures such as area [22], [23] or growth rate [21] (see Table IV (a)). However, in this work, we *generalize* this principle to cover *any* primitive-based measures. Note that when the converter $\underline{c}$ returns no measure (e.g., $bond$ or $aconf$), it means that the measure decreases with respect to the specialization. Dually, $\overline{c}(m) = \emptyset$ means that m increases with respect to the specialization.

In practice, as the skypatterns are computed for a set of measures, we extend the minimal and maximal converters:

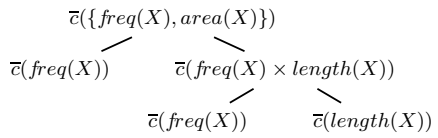
Definition 5 (Minimal and maximal skylineable converters): The minimal and maximal skylineable converters defined by Table III for any primitive-based measure are naturally extended to a set of primitive-based measures $M \subseteq \mathcal{M}$: $\overline{c}(M) = \bigcup_{m \in M} \overline{c}(m)$ and $\underline{c}(M) = \bigcup_{m \in M} \underline{c}(m)$.

Table IV: Applying the minimal and maximal converters

(a) Individual measures

Meas. m	$\overline{c}(m)$	$\underline{c}(m)$
$area$	$\{freq(X)\}$	$\{length(X)\}$
$mean$	$\{min(X.val)\}$	$\{max(X.val)\}$
$bond$	$\{freq(X), freq_\vee(X)\}$	$\emptyset$
$aconf$	$\{freq(X), max(X.val)\}$	$\emptyset$
gr_1	$\{freq(X, \mathcal{D}_1)\}$	$\{freq(X, \mathcal{D}_2)\}$

(b) A set of measures $M = \{freq(X), area(X)\}$



For instance, $\overline{c}(\{freq(X), area(X)\}) = \overline{c}(freq(X)) \cup \overline{c}(area(X)) = \{freq(X)\}$ and $\underline{c}(\{freq(X), area(X)\}) = \underline{c}(freq(X)) \cup \underline{c}(area(X)) = \{length(X)\}$. $\overline{c}(\{freq(X), area(X)\}) = \{freq(X)\}$ means that the most specific patterns (when the frequency remains unchanged) maximizes the measures $\{freq(X), area(X)\}$. The following property formalizes this observation:

Property 3: A set of primitive-based measures $M \subseteq \mathcal{M}$ is minimally $\underline{c}(M)$-skylineable and maximally $\overline{c}(M)$-skylineable.

In our implementation, the user specified set of measures M is parsed through a syntax tree. Following this step, the minimal and maximal skylineable converters are recursively applied to automically compute $\underline{c}(M)$ and $\overline{c}(M)$ (an example is provided in table IV (b) for $M = \{freq(X), area(X)\}$). This process is illustrated in Figure 2 with the edge labelled 1. From now on, the set of measures M' refers to $\underline{c}(M)$ or $\overline{c}(M)$.

C. Distinct and indistinct operators

In the previous paragraphs, we remarked the fact that some skypatterns share exactly the same values on the whole set of measures M' (e.g. $B =_{\{freq\}} AB$). This observation leads to the following question: *Is it possible to find some representatives for a group of indistinct skypatterns?* We show that the answer is yes and that instead of directly evaluating the skypattern query on $\mathcal{L}$, we can compute the skypatterns on a condensed representation of $\mathcal{L}$ and then regenerate the entire set of skypatterns. For this end, we introduce the *distinct operator* which produces condensed representations adequate to M:

Definition 6 (Distinct operator): Given a set of measures $M' \subseteq \mathcal{M}$, the distinct operation for $P \subseteq \mathcal{L}$ with respect to M' and $\theta \in \{\subset, \supset\}$ returns all the patterns X of P such that their generalizations (or specializations) are distinct from X

with respect to M':

$$\mathcal{D}is_\theta(P, M') = \{X \in P | \forall Y \theta X : X \neq_{M'} Y\}$$

where $\theta \in \{\subset, \supset\}$.

Given a set of measures M', the set of free (respectively closed) patterns adequate to M' corresponds exactly to $\mathcal{D}is_\subset(\mathcal{L}, M')$ (respectively $\mathcal{D}is_\supset(\mathcal{L}, M')$). For instance, from our toy example, $\mathcal{D}is_\subset(\mathcal{L}, \{freq\}) = \{A, B, C, D, E, F, AD, AE, BC, BD, BE, CD, CE, DE\}$ and $\mathcal{D}is_\supset(\mathcal{L}, \{freq\}) = \{A, D, E, AB, AC, ABCDEF\}$.

We now introduce the *indistinct operator* that enables the retrieval of all the indistinct patterns from their representatives:

Definition 7 (Indistinct operator): Given a set of measures $M' \subseteq \mathcal{M}$, the indistinct operation returns all the patterns of $\mathcal{L}$ being indistinct with respect to M' with at least one pattern in P.

$$\mathcal{I}nd(\mathcal{L}, M', P) = \{X \in \mathcal{L} | \exists Y \in P : X =_{M'} Y\}$$

For instance, from Table I, the set of patterns that have exactly the same frequency as patterns B and C is $\mathcal{I}nd(\mathcal{L}, \{freq\}, \{AB, AC\}) = \{B, C, AB, AC\}$.

Property 4: Given a set of preserving functions M', one has the following relation for any $P \subseteq \mathcal{L}$ and $\theta \in \{\subset, \supset\}$:

$$\mathcal{I}nd(P, M', \mathcal{D}is_\theta(P, M')) = P$$

In other words, the indistinct operator is the inverse function for the distinct operator. For instance, $\mathcal{I}nd(\mathcal{L}, \{freq\}, \mathcal{D}is_\supset(\{B, C, AB, AC\}, \{freq\})) = \{B, C, AB, AC\}$.

D. Aetheris: Evaluating skypattern query based on skylineability

To compute skypatterns, we would like to confront distinct patterns together instead of individually comparing each pattern. Indeed, the computation of skypatterns with respect to $M = \{freq, area\}$ can be limited to $\mathcal{D}is_\supset(\mathcal{L}, \{freq\})$ because maximal $\{freq\}$-skylineability ensures us that the other patterns are not dominant patterns. For instance, as $AB =_{freq} B$, the $\{freq\}$-skylineability of M gives $AB \succ_M B$ and B cannot be a skypattern. More formally, we know that $\mathcal{S}ky(\mathcal{I}nd(\mathcal{L}, M', \mathcal{D}is_\theta(\mathcal{L}, M')), M) = \mathcal{S}ky(\mathcal{L}, M)$ from Property 4. Theorem 1 now proves that the skypattern operator can be pushed into the indistinct operator:

Theorem 1 (Operational equivalence): If a set of measures M is M'-skylineable with respect to $\theta \in \{\subset, \supset\}$ and M' is a set of measures, then one has:

$$\mathcal{S}ky(\mathcal{L}, M) = \mathcal{I}nd(\mathcal{L}, M, \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M))$$

Proof: Let M be a set of measures M'-skylineable with $\theta \in \{\subset, \supset\}$.

1.$\mathcal{S}ky(\mathcal{L}, M) \supseteq \mathcal{I}nd(\mathcal{L}, M, \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M))$. Let $X \in \mathcal{I}nd(\mathcal{L}, M, \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M))$ and $Y \in \mathcal{L}$. There exist $X' \in \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M)$ such that $X' =_M X$ and $Y' \in \mathcal{D}is_\theta(\mathcal{L}, M')$ such that $Y' =_{M'} Y$ and $Y' \succeq_M Y$ (i.e., M'-skylineability). As X' belongs to $\mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M)$, it cannot be dominated by any pattern of $\mathcal{D}is_\theta(\mathcal{L}, M')$: $Y' \not\succ_M X$. Thus, X is not dominated by Y (i.e., X is a skyline of $\mathcal{L}$ with respect to M) because $X' =_M X$ and $Y' \succeq_M Y$.

2.$\mathcal{S}ky(\mathcal{L}, M) \subseteq \mathcal{I}nd(\mathcal{L}, M, \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M))$. Let $Y \in \mathcal{S}ky(\mathcal{L}, M)$. There exists $Y' \in \mathcal{D}is_\theta(\mathcal{L}, M')$ such that $Y' =_{M'} Y$ and $Y' \succeq_M Y$. As Y is a skypattern, one has $Y \succeq_M Y'$ and thus, $Y' =_M Y$. Furthermore, no pattern of $\mathcal{D}is_\theta(\mathcal{L}, M')$ dominates Y nor Y': $Y' \in \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M)$. Finally, as $Y' =_M Y$, Y belongs to $\mathcal{I}nd(\mathcal{L}, M, \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M))$. ■

It is well-known that the size of adequate condensed representations (i.e., $\mathcal{D}is_\subset(\mathcal{L}, M')$ or $\mathcal{D}is_\supset(\mathcal{L}, M')$) is smaller than the whole collection of patterns [16]. Thus, we have achieved our objective as mentioned in Section III-B. Furthermore, note that if a set of measures is *strictly* M'-skylineable, Theorem 1 reduces to the following relation: $\mathcal{S}ky(\mathcal{L}, M) = \mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M)$ (with $\theta \in \{\subset, \supset\}$). Even if a set of measures is *not* strictly M'-skylineable, it is often preferable not to perform the indistinct operation as done in our case study (see Section V-B). In such situation, the skypatterns of $\mathcal{S}ky(\mathcal{D}is_\theta(\mathcal{L}, M'), M)$ form a condensed representation of $\mathcal{S}ky(\mathcal{L}, M)$.

Figure 3 illustrates the computation of the skypatterns with our approach Aetheris. Suppose that $M = \{freq, area\}$, the first step applies the maximal skylineable converter on M. Then, the distinct operator preserves the closed itemsets (Step 2). The skyline operator selects the dominant patterns at Step 3 by removing D and E which are dominated by AB (i.e., $area(D) = area(E) = 3 < area(AB) = 6$). Finally, the last step computes the indistinct patterns of skypatterns. Note that this step is unnecessary here because the area is strictly $\{freq\}$-skylineable.

E. Discussion

As aforementioned, with itemset patterns and the frequency measure, the distinct operator corresponds to the well-known notions of closed or free frequent pattern condensed representations. Indeed, $\mathcal{D}is_\subset(\mathcal{L}, \{freq\})$ is analogous to free frequent itemsets and $\mathcal{D}is_\supset(\mathcal{L}, \{freq\})$ corresponds to closed frequent itemsets. The pattern mining community provides many efficient algorithms to extract these concise representations. In addition, different studies extend the notion of concise representations to any frequency-based measures or condensable function [25]. These theoretical and algorithmic works support our claim that discovery of skypatterns is very efficient, but also extendable to a very large set of measures. This measure genericity allows the end-user to analyze patterns through multiple and useful criteria.

Evaluating efficiently the distinct operator on more complex patterns such as sequences, trees and graphs implies

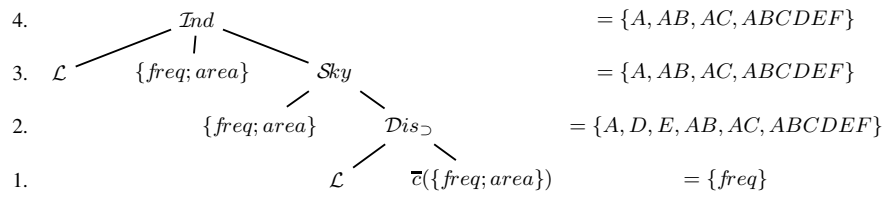


Figure 3: Computing the skypatterns with respect to $\{freq; area\}$ from running example

additional challenges. To cite one example, in the case of sequences, convenient properties such as the *free patterns apriori property* [26], which implies effective search space pruning, cannot be used. Furthermore, in the case of complex patterns, and to the best of our knowledge, no work focused on building concise representations except on the frequency-based measures.

However, it is worth mentioning that Theorem 1 holds for *any* set of measures and *any* language. This means that the efficient extraction of complex skyline patterns (i.e., skyline sequential patterns or skyline graph patterns) is strongly correlated to the advances and progress on complex pattern condensed representations. Last, it is important to notice that Aetheris is not an *exclusive* approach in the sense that it can be coupled with other efficient approaches [27], [28] to extract statistically significant skypatterns.

V. Experimental Study

We report an experimental study on several benchmarks and a case study from chemoinformatics.

A. Experiments on UCI benchmarks

Protocol. Our approach is the first to mine the whole set of skypatterns in a generic way. As a result, we cannot compare it with earlier methods. Nevertheless, for some data sets, skypatterns can be extracted by applying the skyline operator $\mathcal{S}ky$ as a post-treatment on the collection of itemsets that occurs at least once in the dataset, denoted by $\mathcal{L}$. We call this process the **baseline approach**. Our first batch of experiments focus on comparing runtimes of the baseline approach with respect to **Aetheris**. In our experiments, we limit the set of measures M' to preserving functions only. In this way, we can use any mining algorithm adequate to free and closed itemsets [25]. For a fair comparison, the two approaches use the same implementation of the operator $\mathcal{S}ky$ which is based on the block nested loop (BNL) algorithm [9]. Our second batch of experiments aims at comparing our approach to an optimal constraint-based mining method (with thresholds). For each measure $M_i \in M$, we set the threshold σ_{M_i} to $\min_{s \in \mathcal{S}ky\ (\mathcal{L},\ \mathrm{M})}(M_i(s))$. This condition guarantees that no skypatterns will be missed. For instance, in our running example (Figure 1), $\sigma_{freq} = 2$ and $\sigma_{length} = 1$. The set of resulting patterns is called the **optimal constraint-based patterns** (or OCB patterns). This set of patterns needs to be post-processed to find the complete set of skypatterns $\mathcal{S}ky\ (\mathcal{L},\ M)$. Even if this method may seem unrealistic (the user needs to guess optimal thresholds), we still think that this experiment has the benefit of quantifying the reduction of patterns brought by Aetheris even in the scenario where an *ideal end-user* is able to perfectly manage theresholds selections in the constraint-based paradigm.

Datasets and measures. Experiments were carried out on 16 various (in terms of dimensions and density) benchmarks from the UCI repository[2]. We considered a number of combinations of primitive-based measures: frequency, area, maximum, minimum, growth rate and mean. Measures using numeric values were applied on attribute values that were randomly generated within the range [0,1] (see Table I). All the tests were performed on a 2.5 GHz Xeon processor with Linux operating system and 2 GB of RAM memory. Running times were averaged over 5 executions.

Results. Table V and VI provide an overview of 128 experiments carried out on 16 benchmarks, by aggregating the results for 8 sets of measures. Table V presents averages and maximal results for Aetheris and the baseline approach. Note that runtimes only consider the application of skyline operator and do not take into account mining runtimes to extract collection of itemsets (baseline approach) or the pattern condensed representation (Aetheris approach). Mining condensed representations is generally much more efficient than extracting all itemsets [16]. This means that in practice, the gain of Aetheris on the whole process is even much higher than what is reported. However, because the efficiency of the condensed representations is a well-known result in literature, we prefer in these experiments to focus only on the impact of the skyline operator. It should be noted that in some cases the enumeration of all the itemsets fails (e.g., with `mushroom` and `sick` data sets, see [25] for more details). It means that the baseline approach cannot be applied whereas our approach provides the proper set of skypatterns. This point is a major benefit of our approach.

An important result is that Aetheris always outperforms the baseline approach with at least a factor of 10. The distinct operator used to compute skypatterns speeds up the mining in all cases. The reason is that it drastically reduces the size of the input considered by the skyline operator. However,

[2]http://www.ics.uci.edu/~mlearn/MLRepository.html

when the number of measures increases, the collection returned by the distinct operator becomes less compact and skypattern mining becomes less efficient. Nevertheless, in our experiments, the skyline computation remains extremely fast: there are only 3 experiments requiring more than 1 second with the Aetheris approach (experiments with $M = \{freq; max; area; mean\}$ on `austral`, `crx` and `hepatic`) whereas 61 out of the 128 experiments exceed 1 second for the baseline approach.

Figure 4 (a) depicts the performance of the skyline operator for each of the 128 experiments according to the baseline and Aetheris approaches. As expected, the running time of $\mathcal{S}ky$ increases linearly with the number of itemsets in input. The points corresponding to the Aetheris approach are concentrated on the bottom left corner, showing the efficiency of the method.

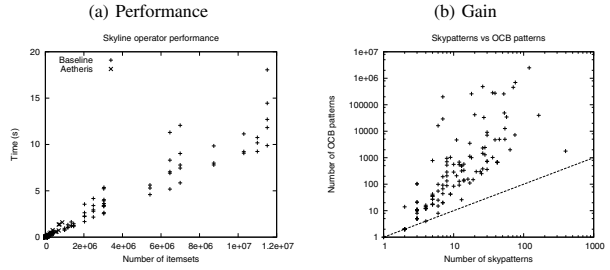


Figure 4: Performance and gain of the skyline patterns.

For each set of measures M, Table VI reports the minimal/average/maximal number of skypatterns, the average number of OCB patterns and the average gain of skypatterns (i.e., |# of OCB patterns|/|$\mathcal{S}ky(\mathcal{L}, M)$|). The aim is to illustrate the problem of *"pattern flooding"* that is still appearing even with the optimal constraint-based approach. In contrast, the number of skypatterns is *always extremely low*. At most, there is a maximum of 397 skypatterns (on `anneal` with the frequency and the growth rate measures). Except for the growth rate measure, a higher number of measures leads to a higher number of skypatterns. The explanation is that a pattern rarely dominates all other patterns on the whole set of measures. Interestingly, the gain of a skyline approach (see the last column in Table VI) is always important (greater than 10 and much greater in almost all the cases). Figure 4 (b) summarizes this result by reporting for each experiment the number of OCB patterns compared to the number of skypatterns. The line $y = x$ highlights the gain of our approach: all the points are above the line and in most cases by several orders of magnitude.

B. Case Study: discovering toxicophores

A necessary step in the elaboration of chemicals' protective measures is the thorough identification of their potentially harmful aspects. Consequently, a major issue in chemoinformatics is to establish relationships between chemicals and a given activity (e.g., LC50 in ecotoxicity). Chemical fragments[3] which cause toxicity are called *toxicophores* and their discovery is a major issue as they are at the core of prediction models in (eco)toxicity [2]. The aim of this case study, which is part of a larger research collaboration with a laboratory of medicinal chemistry, is to investigate the use of skypatterns in order to discover toxicophores.

The dataset is collected from the ECB web site[4]. For each chemical, the chemists associate the data with hazard statement codes (HSC) in 3 acute categories: H400 (very toxic, LC50 $\leq$ 1 mg/L), H401 (toxic, 1 mg/L $<$ LC50 $\leq$ 10 mg/L), and H402 (harmful, 10 mg/L $<$ LC50 $\leq$ 100 mg/L). We focus solely on the H400 and H402 classes. The dataset $\mathcal{D}$ consists of 567 chemicals, 372 from the H400 class and 195 from the H402 class. The chemicals are encoded using 129 frequent subgraphs previously extracted from $\mathcal{D}$[5]. The subgraphs are extracted using a 10% relative frequency threshold (experiments with lower thresholds did not bring significant results for the chemists).

The goal of the first experiment is to evaluate the skypattern approach with measures typically used in contrast mining such as the growth rate since toxicophores are linked to a classification problem with respect to the HSC. When associated together, the growth rate and the frequency measures convey the intuitive notion that a candidate toxicophore is a set of fragments whose frequency is strongly higher in the H400 class than the H402 class and is representative enough (i.e., the higher the frequency, the better it is). We do not specify mining runtimes as they are negligible and we only focus on a qualitative analysis for skypatterns.

A first major result is that the number of skypatterns is very small. Using the growth rate and frequency measures, only 8 skypatterns are enumerated and this allows for a direct expert inspection. The chemists emphasize three patterns based on well-known environmental toxicophores, namely the *phenol ring*, the *chloro-substituted aromatic ring*, and the *organo-phosphorus moiety*. The toxicity of the phenol rings is related to hydrophobocity and formation of free radicals [29]. The chloro-substituted aromatic rings and organo-phosphorus moieties are components of widespread pesticides. Moreover, the organo-phosphorus moiety pattern has a high growth rate (∞ value) and a high frequence. This pattern is thus a jumping emerging pattern and the experts compared it furthermore to jumping emerging fragments (JEF) extracted from previous experiments [30]. It appears that the organo-phosphorus moiety pattern is a generalization

[3]A fragment denominates a connected part of a chemical structure containing at least one chemical bond

[4]ECB, European Chemicals Bureau http://ecb.jrc.ec.europa.eu/documentation/ now http://echa.europa.eu/

[5]A chemical Ch contains an item A if Ch supports A, and A is a frequent subgraph of $\mathcal{D}$.

Table V: Performance analysis of skypattern mining on UCI benchmarks (time in s)

Measures M / θ	Average $\|\mathcal{L}\|$	Average $\|\mathcal{D}is_\theta(\mathcal{L})\|$	Average time base.	Maximal time base.	Average time Aetheris	Maximal time Aetheris	Average gain of Aetheris
$\{freq; area\}$ / maximal (i.e. $\theta = \supset$)	3,754,792.13	63,977.88	3.192	20.110	0.056	0.184	53.82
$\{freq; min\}$ / minimal (i.e. $\theta = \subset$)	3,754,792.13	187,709.69	4.115	26.116	0.194	0.722	18.13
$\{freq; max\}$ / maximal	3,754,792.13	92,459.75	4.150	25.624	0.103	0.396	28.74
$\{freq; max; area\}$ / maximal	3,754,792.13	92,459.75	4.808	29.562	0.122	0.446	28.76
$\{gr; area\}$ / maximal	2,559,789.75	45,489.94	2.280	10.180	0.050	0.176	36.94
$\{freq; gr; area\}$ / maximal	2,559,789.75	45,489.94	2.709	11.146	0.059	0.184	36.97
$\{freq; max; area; mean\}$ / maximal	3,754,792.13	239,017.19	6.361	39.968	0.445	1.600	10.22
$\{freq; gr\}$ / maximal	2,559,789.75	45,489.94	2.274	9.242	0.046	0.144	35.95

Table VI: Effectiveness of skypattern mining on UCI benchmarks

Measures M	Minimal # of skypatterns	Average # of skypatterns	Maximal # of skypatterns	Average # of OCB patterns	Average gain of skypatterns
$\{freq; area\}$	1.00	4.13	8.00	91.81	13.34
$\{freq; min\}$	1.00	4.19	8.00	14403.56	2061.81
$\{freq; max\}$	2.00	10.75	42.00	46748.50	1036.90
$\{freq; max; area\}$	2.00	14.94	57.00	52912.13	1838.87
$\{gr; area\}$	3.00	16.06	71.00	19125.50	1021.52
$\{freq; gr; area\}$	4.00	33.75	75.00	20453.06	399.32
$\{freq; max; area; mean\}$	4.00	35.06	164.00	201596.25	1905.12
$\{freq; gr\}$	6.00	48.44	397.00	2025.94	52.79

of around 90 JEFs and can be seen as a kind of *maximum common structure* (i.e., consensus structure) of these fragments. The experts highly appreciate that Aetheris is able to provide a synthetic view summarizing the information of a large set of JEFs.

The aim of our second experiment is to integrate and evaluate measures conveying a notion of *background knowledge*. In ecotoxicity, chemists consider that the aromaticity and the density measures may yield an interest for candidate toxicophores. For instance, a common hypothesis is that the higher the chemical density, the stronger its chemical behavior. In addition, chemists know that the aromaticity is a chemical property that favors toxicity since their metabolites can lead to very reactive species which can interact with biomacromolecules in a harmful way. Besides, from a biodegradability point of view, aromatic compounds are among the most recalcitrant of the pollutants. Using chemical knowledge, we are able to compute aromaticity and density on chemical fragments. The aromaticity (or the density) of a pattern is calculated using the $mean$ function defined in Table I based on the aromaticity (or density) of each of the 129 listed subgraphs.

Adding only the density to the growth rate and frequency measures do not deeply change the results: 9 skypatterns are obtained and they are similar to the set of 8 skypatterns previously mined with the growth rate and frequency measures. On the contrary, adding the aromaticity and, even better, both the aromaticity and density, leads to skypatterns with novel chemical characteristics. Once again, the whole set of skypatterns remains small (27 when adding the aromaticity and 38 when adding both the aromaticity and the density) and can be directly analyzed by the chemists. They were especially interested in the following skypattern (provided in Smiles code[6]): {`Clc(ccc)c, cc, ccc, cccc, ccccc, ccc(cc)N`}. This skypattern, including an amine function, was not detected during the first experiment and can be exemplified by the chloroaniline derivatives. Indeed, these derivatives are environmentally hazardous since they are very toxic for aquatic species [31]. The experiment shows that background knowledge can successfully be translated to preferences and that Aetheris is straightforwardly able to discover few and promising patterns.

VI. Conclusion

In this paper, we introduce the skyline pattern mining problem. Our goal is to make the result of pattern mining useful from a *user-preference point of view*. We propose Aetheris, the first approach to mine skypatterns in a generic way (i.e., with set of measures and applications to various pattern domains). Aetheris is threshold-free and only needs, as parameters, the measures and the data set. Our approach is based on the key notion of skylineability that supports efficient skypattern computation thanks to an adequate condensed representation of patterns. Experiments performed on several datasets and a use case from chemoinformatics show the efficiency of Aetheris according to both quantitative and qualitative aspects.

An important direction for future work is to improve even further the performance of the algorithm. An idea that we want to investigate is the assimilation of the skyline operator with a pruning strategy. Indeed, Aetheris still applies the skyline operators on pattern collections that may be still relatively large. Other perspectives lie in the improvement

[6]http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

of adequate condensed representations on more complex patterns (i.e., sequences, graphs and dynamic graphs) which is a timely challenge.

Acknowledgements. The authors thank the CERM Laboratory (University of Caen, France) for providing the chemical data and in particular Alban Lepailleur for his highly valuable comments. The authors thank Bertrand Cuissart and Guillaume Poezevara for their contribution for major steps of this work and very fruitful discussions. This work is partly supported by the ANR (French Research National Agency) funded projects BINGO*2* ANR-07-MDCO-014 and FOSTER ANR-2010-COSI-012-02.

REFERENCES

[1] M. J. Zaki and K. Sequeira, "Data mining in computational biology," in *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Press, 2006, ch. 38, p. 1–26.

[2] J. Auer and J. Bajorath, "Emerging chemical patterns: A new methodology for molecular classification and compound selection," *J. Chem. Inf. Mod*, vol. 46, no. 6, p. 2502–2514, 2006.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large database," in *SIGMOD*, 1993, p. 207–216.

[4] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery," *DMKD*, vol. 1, no. 3, p. 241–258, 1997.

[5] F. Bonchi, F. Giannotti, C. Lucchese, S. Orlando, R. Perego, and R. Trasarti, "A constraint-based querying system for exploratory pattern discovery," *Inf. Syst.*, vol. 34, no. 1, p. 3–27, 2009.

[6] S. Bistarelli and F. Bonchi, "Soft constraint based pattern mining," *Data Knowl. Eng.*, vol. 62, no. 1, p. 118–137, 2007.

[7] Y. Ke, J. Cheng, and J. X. Yu, "Top-k correlative graph mining," in *SIAM DM*, 2009, p. 1038–1049.

[8] J. Wang, J. Han, Y. Lu, and P. Tzvetkov, "TFP: An efficient algorithm for mining top-k frequent closed itemsets," *TKDE*, vol. 17, p. 652–664, 2005.

[9] S. Börzsönyi, D. Kossmann, and K. Stocker, "The skyline operator," in *ICDE*, 2001, p. 421–430.

[10] A. N. Papadopoulos, A. Lyritsis, and Y. Manolopoulos, "Skygraph: an algorithm for important subgraph discovery in relational graphs," *DMKD*, vol. 17, no. 1, p. 57–76, 2008.

[11] J. Matousek, "Computing dominances in e^n," *Inf. Process. Lett.*, vol. 38, no. 5, p. 277–278, 1991.

[12] H. T. Kung, F. Luccio, and F. P. Preparata, "On finding the maxima of a set of vectors," *J. ACM*, vol. 22, no. 4, p. 469–476, 1975.

[13] R. E. Steuer, *Multiple Criteria Optimization: Theory, Computation and Application*. John Wiley, 546 pp, 1986.

[14] F. Pennerath and A. Napoli, "The model of most informative patterns and its application to knowledge extraction from graph databases," in *ECML/PKDD*, 2009, p. 205–220.

[15] R. T. Ng, V. S. Lakshmanan, J. Han, and A. Pang, "Exploratory mining and pruning optimizations of constrained associations rules," in *SIGMOD*, 1998, p. 13–24.

[16] T. Calders, C. Rigotti, and J.-F. Boulicaut, "A survey on condensed representations for frequent sets," in *Constraint-Based Mining and Inductive Databases*. Springer, 2004, p. 64–80.

[17] A. Siebes, J. Vreeken, and M. Van Leeuwen, "Item sets that compress," in *SIAM DM*, 2006.

[18] A. Knobbe and E. Ho, "Pattern teams," in *ECML/PKDD*, 2006, p. 577–584.

[19] L. De Raedt and A. Zimmermann, "Constraint-based pattern set mining," in *SIAM DM*, 2007.

[20] B. Bringmann and A. Zimmermann, "The chosen few: On identifying valuable patterns," in *IEEE ICDM*, 2007, p. 63–72.

[21] G. C. Garriga, P. Kralj, and N. Lavrac, "Closed sets for labeled data," *J. Mach. Learn. Res.*, vol. 9, p. 559–580, 2008.

[22] K.-N. Kontonasios and T. De Bie, "An information-theoretic approach to finding informative noisy tiles in binary databases," in *SIAM DM*, 2010, p. 153–164.

[23] F. Geerts, B. Goethals, and T. Mielikäinen, "Tiling databases," in *Discovery Science*, 2004, p. 278–289.

[24] A. Soulet and B. Crémilleux, "Mining constraint-based patterns using automatic relaxation," *Intell. Data Anal.*, vol. 13, no. 1, p. 109–133, 2009.

[25] A. Soulet and B. Crémilleux, "Adequate condensed representations of patterns," *DMKD*, vol. 17, no. 1, p. 94–110, 2008.

[26] D. Lo, S.-C. Khoo, and J. Li, "Mining and ranking generators of sequential patterns," in *SIAM DM*, 2008, p. 553–564.

[27] N. Tatti, "Probably the best itemsets," in *KDD*, 2010, p. 293–302.

[28] G. I. Webb, "Self-sufficient itemsets: An approach to screening potentially interesting associations between items," *TKDD*, vol. 4, no. 1, 2010.

[29] C. Hansch, S. McCarns, C. Smith, and D. Dodittle, "Comparative qsar evidence for a free-radical mechanism of phenol-induced toxicity," *Chem. Biol. Interact.*, vol. 127, p. 61–72, 2000.

[30] S. Lozano, G. Poezevara, M.-P. Halm, and et al., "Introduction of jumping fragments in combination with QSARs for the assessment of classification in ecotoxicology," *Journal of Chemical Information and Modeling*, vol. 50, no. 8, p. 1330–1339, 2010.

[31] E. Argese, C. Bettiol, F. Agnoli, A. Zambon, M. Mazzola, and A. Ghirardini, "Assessment of chloroaniline toxicity by the submitochondrial particle assay," *Environ. Toxicol. Chem.*, vol. 20, p. 826–832, 2001.